Chapter 2
The Theory-Theory

What may well be the most widely accepted theory about the nature of commonsense mental states is the view Morton has labeled *the theory-theory.*[1] 'Functionalism' and the 'causal' theory are more common labels for the doctrine.[2] But each of those terms has acquired such a daunting array of senses and subcategories that we do better to adopt Morton's less encrusted name. In explaining the view, it proves useful to show how it grew out of earlier discussions in the philosophy of mind, and to this end I will do a bit of historical reconstruction. But let the reader beware! The historical approach, for me, is no more than an expository convenience, and my historical sketch is more caricature than portrait.

1. *From Descartes to David Lewis*

The central problem in the philosophy of mind is to explain what mental states *are,* to say how they fit into our broader conception of nature and its categories. Since Descartes a common answer has been that mental states are, quite literally, *sui generis.* The Cartesian view divides reality into two quite distinct though causally interacting domains. My thoughts, perceptions and the like take place in my mind, a "substance" which has no location in physical space, though it does have a special and intimate relation to my body. This special connection between the mental and the physical proved to be Descartes' Achilles' heel. On the one hand it seems undeniable (though not undenied!) that what I think and feel can have a causal effect on the way my body behaves. If I feel a pain in my toe and decide to move it, it moves. Yet with the advance of the physical and biological sciences, it has become increasingly plausible that the physical world is a closed system. The spectacularly successful physicalist paradigm leaves no room for causal intervention from another domain. The textbook version of the problem this poses for Descartes' dualism is often cast in terms of the neurological events intervening between stimulus and behavior: Step on my toe and some nerves will fire; these will cause others to fire; these still others. Finally, though the firing pattern will be staggeringly complex, efferent nerves will fire, muscles will contract, and my toe will move out from under your foot. All this happens without the assistance or intervention of anything nonphysical. So either pains, decisions, and the like are not located in a nonphysical domain, or they are superfluous in the causation of behavior.

Of course this story about neural firings is for the moment no more than hopeful science fiction, since we have no serious idea about the details. Our confidence that the details can eventually be filled in derives from the success of physicalistic theories in other domains. I confess to a nagging suspicion that our confidence may be misplaced. But that is a long story and a different one.

With the decline of Cartesian dualism, philosophers began looking for a way to locate the mental *within* the physical, identifying mental events with some category of events in the physical world. A natural suggestion, in light of the intimate connection between our mental lives and the goings-on in our brains would be to identify mental events as brain events. But under the influence of the verificationist theory of meaning, that view was passed over for the prima facie less plausible view that mental events are (or are definable in terms of, or are logical constructs out of) behavorial events. This, of course, was the central theme of philosophical behaviorism.

It is worth pondering why theorists touched by the verificationist theory of meaning insisted on the primacy of *behavior* in their account of the mental. Granting, for the sake of argument, that all meaningful expressions must be defined in terms of observables, why not attempt to define mental vocabulary in *neurological* terms? My brain states are, after all, just as observable (in principle) as my behavior. The verificationists' preference for behavior was motivated by two closely related lines of argument. First, mental concepts are common coin, shared by the learned and the unlettered. The simplest of souls knows what it means to say 'the injured man is in pain,' or 'Hitler believed he could conquer the world.' But the meaning of these claims for persons of little learning cannot be cashed out in terms of neurological events, since they may be quite ignorant of the fact that they have a nervous system. The second line of argument focuses on language learning. Whether or not we know we have nervous systems, this information surely played no role in our learning to apply mental vocabulary to ourselves and others. The only observable indications of mental states that were available for use in teaching mentalistic vocabulary were behavorial events. So, it was argued, it must be in terms of these that our mental concepts were constructed.

With the growth of philosophical behaviorism, the problem addressed by philosophers of mind underwent a subtle but significant change. For Descartes, Locke, or Berkeley the questions of interest were ontological ones: What sort of thing (or stuff, or process, or substance) is the mind? What sort of thing is matter? How are they interrelated? The philosophical behaviorist, by contrast, is asking not about the nature of the mind, but rather about the *concept* of mind: What is the *meaning* of our mental terms? What is the correct *analysis* of our mental concepts? The shift from ontological questions to questions of conceptual analysis can easily go unremarked when, as was the case with philosophical behaviorism, an answer to the conceptual question trivially entails an answer to the ontological one. If 'S is in pain' means 'S is disposed to behave in certain ways', then, trivially, pain is a behavorial disposition. But, as we shall soon see, there are ways of answering the conceptual question without saying much of interest about the ontological status of mental states and processes.

Before proceeding with our historical reconstruction, it will be useful to have before us some examples of the sort of analyses attempted by philosophical behaviorists. Here are two.

(1) Hempel, in the widely reprinted essay, "The Logical Analysis of Psychology," urges that the meaning or content of any nonobservational empirical statement is given by the various physical tests we would use to determine whether the statement is true. "The statement itself clearly affirms nothing other than this: all these physical test sentences obtain. . . . The statement, therefore, is nothing but an abbreviated formulation of all these test sentences."[3] For the statement 'Paul has a toothache', Hempel offers the following partial unpacking of the abbreviation:

a. Paul weeps and makes gestures of such and such kinds.
b. At the question "What is the matter?" Paul utters the words "I have a toothache."*

*Not all who followed in the tradition of philosophical behaviorism took themselves to be offering analyses of commonsense concepts or everyday locutions. Some took their job to be the analysis of scientific concepts or perhaps the development of new concepts that pass behavioristic muster and might be used in some current or future science. Hempel must have had this latter project in mind, since he completed his list with

c. Closer examination reveals a decayed tooth with exposed pulp. . . .
d. Paul's blood pressure, digestive processes, the speed of his reactions, show such and such changes.
e. Such and such processes occur in Paul's central nervous system.

It seems, however, that the distinction between analyzing ordinary concepts and constructing new ones was often ignored. Hempel gives no hint that the locution he is analyzing is not part of our ordinary vocabulary.

(2) Carnap, offering an account of belief sentences that will cast a substantial shadow in the pages that follow, suggests this as a first pass:

> John believes that snow is white

can be analyzed as

> John is disposed to respond affirmatively to 'Snow is white' or to some sentence which is L-equivalent to 'Snow is white.'[4]

These behavioristic analyses are typical in both their flavor and their difficulties. Most salient among the latter is that the analyses are just plain *false*. The proposed definition or analysis does not capture the meaning or even the extension of the commonsense locution with which it is paired. To see this, consider Paul and his aching tooth. If the analysis is intended to provide necessary and sufficient conditions for 'Paul has a toothache', it fails miserably, for surely Paul may weep, gesture, and respond as indicated while feeling no pain at all; he is just practicing for his role in a dramatization of Mann's *Buddenbrooks.* Conversely Paul may be in utter agony, though as a matter of pride he does not allow himself to give the least indication of his suffering. John, for his part, may believe that snow is white while having not the least disposition to respond affirmatively to 'Snow is white.' He may, for example, believe that such a response would provoke his interlocuter to mayhem; or he may simply wish to deceive others about his beliefs.

These are, to be sure, specific complaints about specific behavioristic definitions, but they augur a more general problem. It is now widely conceded that it is impossible to define psychological expressions in terms of behavior, unless some further psychological terms occur in the definition. John's belief that snow is white will dispose him to respond affirmatively to 'Snow is white' only if he *understands* English, is paying *attention*, *wants* to let you know what he thinks, *believes* that this can be done by responding affirmatively, and has no other *desire* stronger than his desire to let you know what he thinks and incompatible with it. Put simply, the trouble with behavioristic definitions is that when they are not patently inadequate, they are inevitably circular.[5] As awareness of the difficulty sharpened, a number of writers in the philosophical behaviorist tradition explored ways of weakening the central behaviorist claim. Psychological terms were not to be *defined* in terms of behavior; rather, they were to be related to behavior in some less stringent way.[6] But these attempts at rescue generally foundered on one or the other of a pair of perils. Either the new weakened relationship between psychological terms and behavorial terms was too obscure to be understood, or, when the alleged relationship was clear enough, the circularity problem arose anew. Gradually philosophical behaviorism suffered the death of a thousand failures.

While philosophical behaviorism was struggling with the problem of circularity, similar problems were being uncovered by "operationalists" who were attempting to apply the verificationist doctrine to scientific concepts. There too, the doctrine insisted that to be meaningful an expression must be definable in terms of observables. But once again circularity loomed. Consider gravitational mass. We might try something like the following for an operational definition:

> Two objects are equal in gravitational mass = df
> The objects would balance each other on an analytical balance.

However, like the behavoristic definitions considered earlier, this definition is just mistaken; it fails to capture the concept it is trying to define. The test for equal mass will work only if the objects being tested are not being subjected to differential forces. (Imagine that one is wood, the other iron, and the test is being conducted in a strong magnetic field.) But an operational definition of 'subject to equal force' would in turn require some mention of mass in the definiens.

The reaction to this problem in the philosophy of science was to explore a quite different line for explaining how theoretical terms get their meaning. Rather than being *defined* in terms of observables, it was proposed that theoretical terms might get their meaning simply in virtue of being embedded within an empirical theory. The meaning of the theoretical term lies, so to speak, in its theory specified interconnections with other terms, both observational and theoretical. There is nothing mysterious about a term acquiring meaning by being embedded in a theory. Indeed, as a little tale by David Lewis makes clear, it is the sort of thing that happens all the time. Lewis asks that we imagine ourselves

> assembled in the drawing room of the country house; the detective reconstructs the crime. That is, he proposes a *theory* designed to be the best explanation of the phenomena we have observed: the death of Mr. Body, the blood on the wallpaper, the silence of the dog in the night, . . . and so on. He launches into his story:
>
> > X, Y, and Z conspired to murder Mr. Body. Seventeen years ago in the gold fields of Uganda, X was Mr. Body's partner . . . Last week, X and Z confered in a bar in Reading . . . Tuesday night at 11:17, Y went to the attic to set a time bomb . . . Seventeen minutes later, X met Z in the billiard room and gave him a lead pipe . . . Just when the bomb went off in the

(2) Carnap, offering an account of belief sentences that will cast a substantial shadow in the pages that follow, suggests this as a first pass:

> John believes that snow is white

can be analyzed as

> John is disposed to respond affirmatively to 'Snow is white' or to some sentence which is L-equivalent to 'Snow is white.'[4]

These behavioristic analyses are typical in both their flavor and their difficulties. Most salient among the latter is that the analyses are just plain *false*. The proposed definition or analysis does not capture the meaning or even the extension of the commonsense locution with which it is paired. To see this, consider Paul and his aching tooth. If the analysis is intended to provide necessary and sufficient conditions for 'Paul has a toothache', it fails miserably, for surely Paul may weep, gesture, and respond as indicated while feeling no pain at all; he is just practicing for his role in a dramatization of Mann's *Buddenbrooks*. Conversely Paul may be in utter agony, though as a matter of pride he does not allow himself to give the least indication of his suffering. John, for his part, may believe that snow is white while having not the least disposition to respond affirmatively to 'Snow is white.' He may, for example, believe that such a response would provoke his interlocuter to mayhem; or he may simply wish to deceive others about his beliefs.

These are, to be sure, specific complaints about specific behavioristic definitions, but they augur a more general problem. It is now widely conceded that it is impossible to define psychological expressions in terms of behavior, unless some further psychological terms occur in the definition. John's belief that snow is white will dispose him to respond affirmatively to 'Snow is white' only if he *understands* English, is paying *attention*, *wants* to let you know what he thinks, *believes* that this can be done by responding affirmatively, and has no other *desire* stronger than his desire to let you know what he thinks and incompatible with it. Put simply, the trouble with behavioristic definitions is that when they are not patently inadequate, they are inevitably circular.[5] As awareness of the difficulty sharpened, a number of writers in the philosophical behaviorist tradition explored ways of weakening the central behaviorist claim. Psychological terms were not to be *defined* in terms of behavior; rather, they were to be related to behavior in some less stringent way.[6] But these attempts at rescue generally foundered on one or the other of a pair of perils. Either the new weakened relationship between psychological terms and behavorial terms was too obscure to be understood, or, when the alleged relationship was clear enough, the circularity problem arose anew. Gradually philosophical behaviorism suffered the death of a thousand failures.

While philosophical behaviorism was struggling with the problem of circularity, similar problems were being uncovered by "operationalists" who were attempting to apply the verificationist doctrine to scientific concepts. There too, the doctrine insisted that to be meaningful an expression must be definable in terms of observables. But once again circularity loomed. Consider gravitational mass. We might try something like the following for an operational definition:

> Two objects are equal in gravitational mass = df
> The objects would balance each other on an analytical balance.

However, like the behavoristic definitions considered earlier, this definition is just mistaken; it fails to capture the concept it is trying to define. The test for equal mass will work only if the objects being tested are not being subjected to differential forces. (Imagine that one is wood, the other iron, and the test is being conducted in a strong magnetic field.) But an operational definition of 'subject to equal force' would in turn require some mention of mass in the definiens.

The reaction to this problem in the philosophy of science was to explore a quite different line for explaining how theoretical terms get their meaning. Rather than being *defined* in terms of observables, it was proposed that theoretical terms might get their meaning simply in virtue of being embedded within an empirical theory. The meaning of the theoretical term lies, so to speak, in its theory specified interconnections with other terms, both observational and theoretical. There is nothing mysterious about a term acquiring meaning by being embedded in a theory. Indeed, as a little tale by David Lewis makes clear, it is the sort of thing that happens all the time. Lewis asks that we imagine ourselves

> assembled in the drawing room of the country house; the detective reconstructs the crime. That is, he proposes a *theory* designed to be the best explanation of the phenomena we have observed: the death of Mr. Body, the blood on the wallpaper, the silence of the dog in the night, . . . and so on. He launches into his story:
>
>> X, Y, and Z conspired to murder Mr. Body. Seventeen years ago in the gold fields of Uganda, X was Mr. Body's partner . . . Last week, X and Z confered in a bar in Reading . . . Tuesday night at 11:17, Y went to the attic to set a time bomb . . . Seventeen minutes later, X met Z in the billiard room and gave him a lead pipe . . . Just when the bomb went off in the

> attic, X fired three shots into the study through the French windows . . .
>
> And so it goes: a long story.
>
> The story contains the three names, 'X', 'Y', and 'Z'. The detective uses these new names without explanation, as though we knew what they meant. But we do not. We never used them before, at least not in the senses they bear in the present context. All we know about their meaning is what we gradually gather from the story itself.[7]

Yet by the time the detective is finished with his reconstruction, we know the meaning of 'X', 'Y', and 'Z' as well as the detective himself. We may discuss X's motives, debate whether Y was a drug addict, perhaps even disagree about the plausibility of the alleged meeting in the bar. Lewis gives the label *implicit functional definition* to this process of introducing terms by recounting a theory in which they play a role.

Before we see how all this ties up with mental concepts, four observations about implicit functional definitions are in order. First, as Lewis notes, the story the detective tells about X, Y, and Z would be essentially unchanged if, instead of simply plunging into his narrative, he had prefaced his remarks by saying: "There exist three people whom I shall call 'X', 'Y', and 'Z'." The "theoretical terms" 'X', 'Y', and 'Z' in the story would then function as variables bound by an existential quantifier.

Second, it is possible for an implicit functional definition of a set of terms to be quite noncommittal about various facts concerning the "theoretical entities" it introduces, including facts that may be of considerable importance. In the story at hand the detective did not tell us what we probably most wanted to know: the *names* of the persons involved. Perhaps when the tale is told our background knowledge will enable us to name the perpetrators ourselves. But perhaps not. The detective himself may have no idea of the bomber's (as they say) true identity. It is also possible, though this does not emerge in Lewis's tale, for an implicit functional definition to be noncommittal about the *ontological status* of the theoretical objects defined. Suppose, for example, that our detective is something of a spiritualist. He theorizes that there was something, call it 'W', which caused the French windows to slam shut at the crucial moment. That something, W, witnessed a scene between X and Y, and W became infuriated by what he (it?) saw. In short, W enters into a variety of causal relations with other characters in the story, with various physical objects, and so on. Now you may be convinced that W is a person, since you are convinced that only a person could fill that role. But the detective is not convinced. He is willing to entertain the possibility that W is not a physical object at all. Perhaps W is a spirit, a disembodied Cartesian mind. Note that nothing in his theory—which, recall, *is* the implicit functional definition of 'W'—commits him one way or the other on this matter. The theory simply claims that there is *something* which had various causal interactions, and it is neutral on the issue of the ontological status of that something.

The third observation is a corollary to the second. Since implicit functional definitions leave much unsaid about the theoretical entities denoted by theoretical terms, they leave ample room for further elaboration. We might discover a great deal about X that is unknown to the detective and unmentioned in his theory: that X was born in Rhodesia, perhaps, or that he lost a toe due to frostbite. The point is that by further investigation we might come to know more about the detective's theoretical entities than does the detective himself. In the extreme, though it raises some ticklish questions, we might even come to know that the detective was wrong in some of his beliefs about X. This in any event would be the natural conclusion if we were to discover a man who fit the detective's characterization to a tee, save for the bit about meeting Z in the billiard room. (Actually, it was in the adjacent pantry; but X walked through the billiard room to get to the pantry, and it was while walking throught the billiard room that he lost his cufflink.)

The fourth point is that *various* kinds of relations may play a role in implicit functional definitions. In Lewis's story characters are described in terms of their conversations, their business relations, their spatial locations at certain times, and so forth. One sort of relation that does not enter into Lewis's story but well might is what could be called the relation of *typically causing*: Fido is a placid beast, kind to cats, children, even postmen. But for some reason X typically sets Fido off on a howling fit. Not always, and not only X; occasionally Fido will greet X with a wagging tail, and once in a while just about anything will set Fido off. Yet it is the sight of X which typically causes Fido's fits.[8]

So much for implicit functional definition. Now let us turn to what I am calling the theory-theory, which is an attempt to apply the idea of implicit functional definition in the philosophy of mind. The basic thought is that commonsense mental terms gain their meaning just as 'X', 'Y', and 'Z' did in Lewis's detective story. They are "theoretical terms" embedded in a folk theory which provides an explanation of people's behavior. The folk theory hypothesizes the existence of a number of mental states and specifies some of the causal relations into which mental states enter—relations with other mental states, with environmental stimuli, and with behavior. Of course the theory implicitly defining mental terms is rough and incomplete. It may make ample use of the notion of typical causing, allowing for exceptions to causal regularities which are neither explained nor enumerated by the theory. Lewis suggests that we could assemble the relevant folk theory by simply collecting platitudes:

> Collect all the platitudes you can think of regarding the causal relations of mental states, sensory stimuli, and motor responses. . . . Add also all the platitudes to the effect that one mental state falls under another—'toothache is a kind of pain', and the like. . . . Include only platitudes which are common knowledge among us—everyone knows them, everyone knows that everyone else knows them, and so on. For the meanings of our words are common knowledge, and the names of mental states derive their meaning from these platitudes.[9]

Let us try to make this proposal a bit more precise. A fragment of folk theory about mental states might include a dozen mental predicates (like 'has a toothache', 'is afraid,' 'is thinking of Vienna') which we may abbreviate M_1, M_2, . . . , M_{12}. It will also include some predicates characterizing environmental stimuli, say S_1, S_2, . . . , and some predicates characterizing behavior, say B_1, B_2, ... (The predicates may take any number of arguments, though for simplicity of exposition I will assume they are all monadic.) Some of the generalizations of folk psychology specify typical causal relations among mental states. If we let the arrow '→' mean 'typically causes', then one of these generalizations might be represented as follows:

(1) $(x)\ M_1x\ \&\ M_2x \rightarrow M_5x\ \&\ M_8x.$

Other folk platitudes tell of causal relations between stimuli and mental states, thus:

(2) $(x)\ S_2x\ \&\ S_7x \rightarrow M_{12}x.$

Still others detail causal relations between mental states and behavior:

(3) $(x)\ M_8x\ \&\ M_4x \rightarrow B_6x.$

Finally, still others may detail complex relations with a stimulus and a mental state causing both some behavior *and* some subsequent mental states:

(4) $(x)\ S_3x\ \&\ M_4x \rightarrow M_7x\ \&\ B_{12}x.$

The whole of our folk psychological theory might be represented as a conjunction of principles similar in form to (1)–(4). Or, paralleling a move mentioned in the discussion of Lewis's detective story, we might replace all of the mental *predicates* with variables and preface the long conjunction of folk principles with a string of existential quantifiers. The result would be a Ramsey sentence of the form

(5) $(\exists m_1)(\exists m_2) \ldots (\exists m_k)\ (x)\ T$,

where 'T' is the conjunction of (1)–(4) and similar principles, with variables suitably substituted for mental state terms.

A number of points should be stressed about the theory-theory account of the meaning of commonsense mental terms. First, if the account is correct, then the ordinary use of mental state terms carries a commitment to the *truth* of our folk theory. If the folk theory expressed in a sentence like (5) turns out to be false, then the mental state predicates the theory implicitly defines will be true of nothing. We needn't be purists on the point, insisting that every detail of our folk theory turn out to be true. Various strategies are available for sidestepping this unwelcome implication.[10] But there is no escaping the fact that for a theory-theorist, most of our folk platitudes must be true if any attribution of a mental state is true. As Lewis notes, if the theory-theory is true, then "mental terms stand or fall together. If common-sense psychology fails, all of them alike are denotationless."[11]

My second point is that the theory-theory is ontologically noncommittal. Mental states are characterized by the role they play in a complex causal network purporting to explain behavior. But just as the detective's story may be noncommittal on whether the thing filling the role of W is a person or a disembodied spirit, so too the folk theory which implicitly defines mental state terms remains neutral on whether the fillers of the various causal roles are physical states, Cartesian mental states, or what have you. For the advocate of the theory-theory, this "topic neutrality" is a singular virtue, since it allows what seems prima facie obvious, viz., that a Cartesian and his materialist neurologist can communicate quite successfully using the vocabulary of folk psychology. If, as the theory-theory insists, mental terms are defined by topic neutral folk theory, then despite the differences in their further beliefs, the Cartesian and the neurologist mean the same when they invoke commonsense mental terms.

On the theory-theorist's view, it is a matter for science to determine just what fills the causal roles specified by folk theory. Though there is an important ambiguity in putting the matter this way. Assuming that folk psychology is true (or near true) of both you and me, we may ask exactly what it is that fills the causal role which folk platitudes specify for my current backache or for your current perception of a printed page. And surely the best scientific bet in both cases is that the causal roles are filled by current physical states of our brains—very different states, no doubt, for my backache and your perception. So the theory-theory along with some timid scientific speculation entails the *token identity theory*. Specific, individual, dated occurrences of mental states are to be identified with specific, dated occurrences in the brain of the person having the mental state. But what about mental state *types*? What, in general, are toothaches or thoughts of Vienna? If it were the case that for all toothaches the platitude-characterized causal role is filled by neurologically identical brain states, then it might be plausible to say that a toothache (the type now, not a token) simply *is* a type of neurological event. This is the doctrine known as the *type identity theory*. However, the premise on which the type identity theory rests, viz., that all individual toothaches (tokens) are identical with brain states of the same neurological type, may well turn out to be false. It is a good bet that different humans, or humans and sharks, or humans and Martians, typically have *different* sorts of brain states playing the causal role that folk psychology assigns to a toothache. Considerations of this sort have led most theory-theorists[12] to conclude that a state type like having a toothache is best viewed not as a physical state type, but rather as a "functional" state. On this view, the state (type) of having a toothache (for anyone, or anything, at any time) is the complex state an organism is in when it (1) satisfies the precepts of folk psychology (i.e., is truly described by folk psychology), and (2) is in some (presumably but not necessarily physical) state which, in the person or organism in question, fills the causal role assigned to toothaches by folk theory.

The third point to stress about the theory-theory is one which will take on great importance in subsequent chapters. Theory-theorists typically hold that commonsense psychology is a theory aimed at explaining behavior *in terms of the causal relations among stimuli, hypothesized mental states, and behavior*.[13] To adopt this view of folk psychology is to exclude any reference to noncausal relations in folk psychology. There can be no mention of a subject's social setting, natural environment, or personal history, nor of the psychological characteristics of other people. And since for the theory-theorist mental terms are implicitly defined by commonsense psychological theory, none of these factors can be conceptually tied to commonsense mental terms or concepts.[14] Thus the theory-theory is committed to what I will call the *narrow causal individuation* of mental states. Perhaps the best way to explain this terminology is to focus on the question of what must be true of a pair of subjects if they are to be in mental or psychological states of the same type. What, for example, must be true of Fido and me if we both have a toothache?[15] The answer provided by the theory-

theory goes like this: First, both of us must be truly described by folk theory. Second, each of us must be in a state which fills the role assigned to toothaches by folk theory. Note that for these two conditions to obtain, it is not necessary for Fido and me to be at all similar in our chemical or physiological makeup. Fido could perfectly well be a robot whose brain is made of silicon chips. But it is necessary that some state of Fido's be *causally isomorphic* with a state of mine. That is, his token and mine must exhibit parallel patterns of potential causal interactions as these are characterized by folk theory. When an account of mental states holds that only their respective patterns of causal interactions count in determining whether a pair of states are of the same mental or psychological type, I will say that the account is committed to a *causal individuation* of mental states. Now what about the "narrow"? Well, causal links can extend well beyond the boundary of an organism. There are, no doubt, causal principles of various sorts linking some of my current psychological states to events that occurred long before I was born. But on the theory-theory account of mental states, none of these causal links are relevant in determining the mental or psychological type of the state. The only potential causal links recognized by the theory-theory are those that obtain between mental states and other mental states, those that obtain between mental states and *stimuli* and those that obtain between mental states and *behavior*. Accounts of causal individuation which restrict attention to just these causal links, I will call *narrow* accounts. In chapters 4 and 5 I will argue that this is a fundamental mistake, that no account committed to a narrow causal individuation of mental states can do justice to our folk psychological concept of belief. But before our stance turns critical we should take a more detailed look at the sort of account of belief that the theory-theory can offer.

2. *Belief and the Theory-Theory: Some Problems*

Thus far I have been discussing the theory-theory as a general account of the meaning of mental terms. In the current section I want to explore the prospects of applying this account to belief ascriptions—ascriptions of the form 'S believes that p'. In setting out the theory-theory, a commonly used illustration is pain, or some specific sort of pain, like a toothache. But there are a number of important differences between mental states like having a toothache, and mental states like believing that Socrates was wise, differences which make it hard to see how the theory-theory can provide an account of belief at all. First, pain has relatively strong and direct links to both environmental stimuli and to behavior; the links between belief and stimuli or behavior are much more tenuous. Second, folk psychology specifies relatively few links between pains and other mental states; beliefs, by contrast bristle with causal links to other mental states (including other beliefs). Third, just about all the varieties of pain for which our folk vocabulary provides labels have had many instances in the actual world. By contrast, folk psychology provides standard designations for an endless variety of beliefs that no one has ever held. Finally, pain is not prima facie a relational notion; 'has a toothache' is a one-place predicate. But belief at least *appears* to be a relational notion. It is tempting to say 'John believes Socrates was wise' expresses a *relation* between John and something, though, notoriously, it is not easy to say just what the second element of the relation is. These four differences will serve as convenient foci in the discussion to follow.

To see how the first of these differences makes belief problematic for the theory-theorist, consider the sort of account that could be given of a mental state like having a pricking pain in the finger. What folk platitudes might contribute to an implicit functional definition of this notion? Well, pricking pains in the finger are typically caused when the skin on the finger is pricked with a pin or other sharp object. In turn, pricking pains typically cause the finger to be quickly withdrawn. Pricking pains often cause wincing and a variety of verbal exclamations, many of which are not printable. They may also cause brief rubbing of the site of the wound. A few more platitudes might be added, but we already have a good sampling of the folk wisdom about pricking pains in the finger. On the theory-theorist's account, the pricking pain is defined as the state which has these characteristic causes and effects. If we try to tell a parallel story for beliefs, however, problems quickly crowd in on us. For, in contrast to pain, there is generally no characteristic environmental stimulus which typically causes a belief. There is no bit of sensory stimulation which typically causes, say, the belief that the economy is in bad shape, or the belief that Mozart was a Freemason. Beliefs about one's current surroundings are arguably an exception here; the belief that one is standing before an elephant *is* typically caused by having an elephant before one's eyes. But the fact remains that the bulk of our beliefs have no typical or characteristic environmental causes.* Nor do most beliefs have typical behavioral effects. My belief that Ouagadougou is the capital of Upper Volta does not cause me to do much of anything. It does (or would) typically cause

*This is not, of course, to say that these beliefs do not sometimes *have* environmental causes, but only that they do not have typical or characteristic causes of the sort needed for the theory-theorist's implicit definition. Your belief that the economy is in bad shape may have been caused by the sight of a price sign at your local gas station; mine was caused by the sight of the letter detailing my laughable pay increase.

me to say 'Ouagadougou is the capital of Upper Volta' in response to 'Fifty bucks if you can name the capital of a central African nation.' But it certainly does not cause most people who share my belief to utter what I would, given the same stimulus. For most of the people who share my belief, including most of the inhabitants of Upper Volta, neither speak nor understand English.

Since beliefs generally do not have typical stimulus causes or typical behaviorial effects, the theory-theorist must look elsewhere for platitudes to incorporate in implicit functional definitions. With links to stimuli and to behavior generally unavailable, the only alternative open to the theory-theorist is to build implicit functional definitions from principles detailing the relations of beliefs to other beliefs and to mental states of other sorts. This path is strewn with obstacles, however, and in what remains of this chapter, I will survey some of them.*

The most salient and systematic of the causal relations linking beliefs with other mental states are those that fall under the headings *inference* and *practical reasoning*. Here are some examples: Typically, if a person believes that everyone aboard the morning flight to Bamako was killed in a crash, and if he comes to believe that the prime minister of Mali was on that flight, those two beliefs together will cause him to believe that the prime minister is dead. Typically, if a person wants to go to New Zealand, and if she comes to believe that she must get a visa in order to make the trip, she will acquire a desire to get a visa. Endlessly many further examples might easily be constructed. But therein lies a problem, for there is no hyperbole in the prospect of *endlessly many* further examples. There are literally infinitely many inferential paths leading both to and from every belief. Moreover a description of each of these inferences (or at least of an infinite subset of them) is arguably part of our shared fund of folk platitudes about belief. An example may help to make the point obvious. Consider the inferences that would typically be drawn from the belief that Niger is west of Mali. If a person believes this and comes to believe that if Niger is west of Mali, then Niger has more than one million inhabitants, he will typically come to believe that Niger has more than one million inhabitants. If a person believes that Niger is west of Mali and comes to believe that if Niger is west of Mali then Niger has more than two million inhabitants, he will typically come to believe that Niger has more than two million inhabitants. And so on, ad infinitum. Moreover, obviously, this is only

*At this point we part company with Lewis who, as far as I know, has not addressed the problem of providing implicit functional definitions for belief predicates. The only serious and detailed attempts to provide a causal role analysis of belief that I know of are in Armstrong 1968 and 1973. For a critique of Armstrong's analysis, see Stich 1983.

the beginning. (Suppose he believes that if Niger is west of Mali, then Niger grows 5 percent more grain, or that if Niger is west of Mali, then Mali is more politically stable, or . . .) If we are to construct our implicit functional definition of 'believes that Niger is west of Mali' from the conjunction of these commonsense platitudes about inference, then the definition will have to be *infinitely long*! Now some stout-hearted souls are not intimidated by infinitely long conjunctions, provided they can be recursively summarized.[16] But when we reckon in inductive inference as well as deductive inference, it is far from clear that the class of inferences recognized by common sense *is* recursively characterizable. Even if we assume it is, there is considerable awkwardness in the view that the meaning of workaday belief attributions can be specified only by an infinite conjunction. Thus we have some motive to hunt for an alternative account.

The problem we have been discussing derives from the fact that common sense is so rich in platitudes about belief. But there is also a sense in which common sense is not rich enough in such platitudes, and this generates another sort of problem for the strategy of analyzing belief ascriptions in terms of intramental causal links. While there are infinitely many inferences with a claim to being embedded in common sense, there are also infinitely many logically valid inferences which people neither draw nor expect others to draw. If we know that Maggie believes all Italians love pasta, and if we know she has just learned that Sven is an Italian, we expect Maggie will come to believe that Sven loves pasta. But suppose that instead of believing Sven is Italian, Maggie comes to believe some claim which is logically equivalent, though it would take a clever logician a week of hard work to prove the equivalence. Plainly, we would not, then, expect Maggie would come to believe that Sven loves pasta. All this is unproblematic. Trouble arises because the line between cases of the first sort and cases of the second sort is anything but sharp. There are enormous numbers of inferences which some people expect their fellows to draw, while others do not. Which of these are we to take as part of the body of platitudes which define belief predicates? Advocates of the theory-theory give little guidance here, and I suspect that any answer will be ad hoc and implausible.

Another problem for the intramental definition strategy emerges when we reflect on the motive for invoking the notions of *typical cause* and *typical effect* in theory-theoretic accounts. Briefly, the story goes like this. Folk psychology recognizes that the causal regularities from which it is woven do not always obtain. For a given stimulus to produce a given mental state or for one mental state to produce another, conditions have to be suitable. Some of the ways in which conditions can be suitable or unsuitable are incorporated into further folk principles. But many are not. This causes no problems so long as the causal links folk theory specifies generally obtain under familiar circumstances. Thus a pricking pain in the finger can be characterized as the state typically caused by a pin pricking the finger, etc., and this characterization will pick out the right state, despite the occasional instance of a finger pricking causing no pain or of a pricking pain in the finger caused by a blow on the head. Note, in all this, that the notion of typicality being invoked is an utterly pedestrian one. A typical effect is just an effect that arises *in a substantial majority of the relevant cases.*

When we try to transfer this notion to the case of belief, however, things get very messy, for there are just not enough actual beliefs to go around. To jaded academic ears this may sound like an odd claim, since it often seems (as a colleague once put it to me) that every belief, no matter how outrageous, has at least a few adherents. But surely this is more pessimism than the facts warrant. No one believes—no one has *ever* believed—that Buckingham Palace is filled with pickles from floor to ceiling. And a few minutes of playful thought will be enough to convince you that there are endlessly many other beliefs that no one has ever held. Once this is granted, the strategy of defining beliefs by specifying their *typical* causes and effects runs into trouble. In the case of pain, 'typical' was unpacked as 'true of a substantial majority.' But what about such predicates as 'believes that Buckingham Palace is filled with pickles'? Since no one has ever held this belief, it has never caused or been caused by anything. What then are we to make of talk about its *typical* causes and effects? In chapter 5 we will see that an analogous problem arises for the account of belief ascription I want to defend. The solution I sketch there, however, will be of no help to the theory-theorist pursuing the intramental link strategy on beliefs.

A final difficulty with the strategy we have been considering is that it entails what Field has called the "orthographic accident view."[17] There is, as we noted earlier, a very strong temptation to insist that a statement like 'Jack believes that Bamako is the capital of Mali' expresses a *relation*. Jack is said to stand in some relation to a proposition, or fact, or sentence about Bamako. 'Jack believes that Niamey is the capital of Mali' asserts that Jack stands in the same relation to a different proposition or sentence. But on the view I have been sketching, predicates like 'believes that Bamako is the capital of Mali' are treated as one place predicates, whose meanings are implicit in a theory specifying the potential causal antecedents and consequences of the state denoted by the predicate. On this view it is simply an accident of nomenclature that 'believes' occurs in both 'believes that Bamako is the capital of

Mali' and in 'believes that Niamey is the capital of Mali,' an accident of no more significance than the fact that 'It' occurs in both 'Italy' and 'It's raining.' It is also an orthographic accident that 'Mali' occurs in both belief predicates. Indeed, it seems to follow from this view that beliefs are not (or at least need not be) a natural folk psychological category at all. The states attributed by 'believes that Socrates was wise', 'believes that Frege was mistaken', and 'wants to visit New Zealand' are each characterized by a set of typical causes and effects specified by folk theory. There is nothing to indicate that the first two are states of the same kind, different from the third. To make matters worse, the orthographic accident view makes it something of a mystery how we ever succeeded in mastering the language of belief ascription. We all understand infinitely many predicates of the form 'believes that p', though on the orthographic accident view each of these is a logically atomic predicate like 'is red'.

All of this is wildly implausible.* Surely it is more promising to view belief sentences as expressing a relation of some sort. And, conveniently enough, this thought leads directly to the *mental sentence* theory of belief, which is the topic of the next chapter. But what shall we conclude about the theory-theory account of belief? Well, at a minimum it is clear that the view must confront some very substantial obstacles. Perhaps enough has already been said to convince you that the theory-theory strategy cannot produce an account which does justice to our folk notion of belief. If not, then read further. In chapters 4 and 5 an argument will be developed which applies, mutatis mutandis, against any account of belief which individuates beliefs on narrow causal lines. Since the theory-theory is committed to narrow causal individuation, that argument gives yet another reason to conclude that the theory-theory tells the wrong story about the commonsense concept of belief.

*Though not so implausible that it lacks eminent advocates. Quine 1960 comes very close to adopting the orthographic accident view.